SANDIA REPORT
SAND 2011-8448
Unlimited Release
October 2011

Robust Automated Knowledge Capture

Robert G. Abbott, Michael Haass, Michael Trumbo, Susan Stevens-Adams, Stacey Hendrickson & Chris Forsythe

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND 2011-8448
Unlimited Release
October 2011

Robust Automated Knowledge Capture

Robert Abbott & Michael Haass
Cognitive Systems

Michael Trumbo
Cognitive Modeling

Susan Stevens-Adams & Stacey Hendrickson
Human Factors and Statistics

Chris Forsythe
Cognitive Modeling

Sandia National Laboratories
P.O. Box 5800
Albuquerque, New Mexico 87185-1188

Abstract

This report summarizes research conducted through the Sandia National Laboratories Robust Automated Knowledge Capture Laboratory Directed Research and Development project. The objective of this project was to advance scientific understanding of the influence of individual cognitive attributes on decision making. The project has developed a quantitative model known as RumRunner that has proven effective in predicting the propensity of an individual to shift strategies on the basis of task and experience related parameters. Three separate studies are described which have validated the basic RumRunner model. This work provides a basis for better understanding human decision making in high consequent national security applications, and in particular, the individual characteristics that underlie adaptive thinking.

ACKNOWLEDGMENTS

Substantial portions of the work reported here were conducted by researchers at the University of Notre Dame under the leadership of Dr Gabriel Radvansky and the University of Memphis under the leadership of Dr Arthur Graesser and Dr Sidney D'Mello through contract to Sandia National Laboratories. Sections of this report have been excerpted from progress reports provided to Sandia National Laboratories by these research teams.

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

CONTENTS

FIGURES

TABLES

NOMENCLATURE

ACT	American College Test
AMT	Amazon Mechanical Turk
HIT	Human Intelligence Tasks
LDRD	Laboratory Directed Research and Development
OSPAN	Automated Operation Span Task
MAT-B	Multi-Attribute Test Battery
RAKC	Robust Automated Knowledge Capture
RAT	Random Associates Task
SAT	Scholastic Aptitude Test
SNL	Sandia National Laboratories

1. INTRODUCTION

In formulating the Robust Automated Knowledge Capture LDRD project, the objective had been to establish relationships between individual aptitudes on different measures of cognitive performance and an individual's predilection to select different strategies for a given task. This research question was based on a dynamical systems theory perspective regarding brain function (Port & Van Gelder, 1995). In accordance with this theory, the brain consists of tissue specialized for performing various computational algorithms and is in essence a collection of these algorithms. Different individuals possess varying endowments with respect to their capacity to execute various algorithms with these differences reflected in various measures of cognitive aptitude. Within everyday settings, as an individual is faced with various cognitive demands, they bring together combinations of algorithms that manifest as a specific strategy. Given individual differences in cognitive aptitudes, different individuals may bring together different combinations of algorithms with the product being somewhat different strategies. For example, if asked to memorize a list of words, an individual proficient with mental imagery may evoke mental images of each item whereas an individual proficient with verbal processes may construct sentences or merely repeat the word. Consequently, a key hypothesis was that ***for a given task, individuals will exhibit different strategies with the specific strategy employed being a product of their intrinsic skills***.

The second component of a dynamical systems perspective considers the role of feedback. It is assumed that there will be feedback regarding the effectiveness of actions taken in response to various cognitive demands. However, individuals should differ with respect to their capacity to utilize positive and negative feedback in evaluating selected strategies and switching from less effective to more effective strategies, a trait that has been referred to as adaptability. Thus the second key hypothesis was that ***individuals will exhibit varying levels of adaptability with an individual's adaptability determining their propensity to switch strategies in response to changing conditions***.

The following report summarizes experiments conducted over the course of the three-year project to test and expand upon the above hypotheses.

2. FIGURE-8 DRAWING TASKS

In this study, a task was employed, the Figure-8 Drawing Task, which required subjects to trace a figure-8 depicted on a touch tablet. Although seemingly simple, this task offered several desirable characteristics. The horizontal and vertical symmetry allow for consistency when modifying the stimulus or instructions and analyzing the data. Moreover, it is a moderately complex task, thereby negating a potential lack of understanding of the task, which could inhibit optimal strategy selection. Such a task is also largely free of variations in prior knowledge or expertise (as opposed to chess, for example), yet allows for multiple strategies.

Subjects participated in 2 two-hour sessions. During the first hour, they performed the figure-8 drawing task and during the second hour, they completed a battery of cognitive aptitude measures. With this task, as shown in Figure 1, a figure was presented on a Wacom Techno Cintiq 21UXTM system utilizing a 43 cm x 33 cm LCD monitor. Using a stylus, subjects completed a series of experimental conditions that varied with respect to the drawing conditions and instructions given to the subjects. As illustrated in Figure 2, after each trial, subjects were provided a feedback score, which depending on the experimental condition, involved a weighted composite of their speed and accuracy.

Figure 1 Apparatus used for Figure-8 Drawing task.

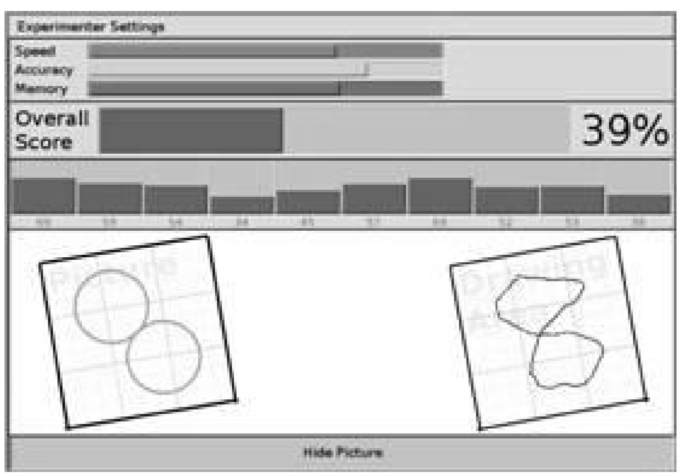


Figure 2 Feedback was presented on display monitor following each trial.

Six experimental conditions were tested which included:

(1) *practice*, which emphasized familiarization with the system with there being no performance feedback;
(2) *baseline*, which included feedback in which speed and accuracy were equally weighted (see Figure 3);
(3) *random size*, which varied the size of the figure subjects were asked to trace (see Figure 3);
(4) *infinity*, with figure-8 presented horizontally and described as an infinity symbol (see Figure 3);
(5) *memory*, in which different images (e.g. five-pointed star, SNL logo) were briefly presented and the subject was asked to draw the image from memory;
(6) *angle trace*, in which figure-8 was rotated at various angles (see Figure 3);
(7) *angle draw*, in which figure-8 was presented at different angles and the subject was asked to reproduce the figure in an adjacent panel, as opposed to tracing the figure;
(8) *speed*, subjects were asked to focus on speed and the feedback was adjusted accordingly;
(9) *accuracy*, subjects were asked to focus on accuracy and the feedback was adjusted accordingly;
(10) *no ink*, subjects used the stylus to trace the image, but the pen produced no marks, and thus, provided no immediate feedback;
(11) *Random interstimulus interval (ISI)*, in which the interval between trials was randomly varied; and
(12) *Trial timeout*, in which after a random duration had passed, if the subject was not complete, the trial would timeout giving a composite score of zero.

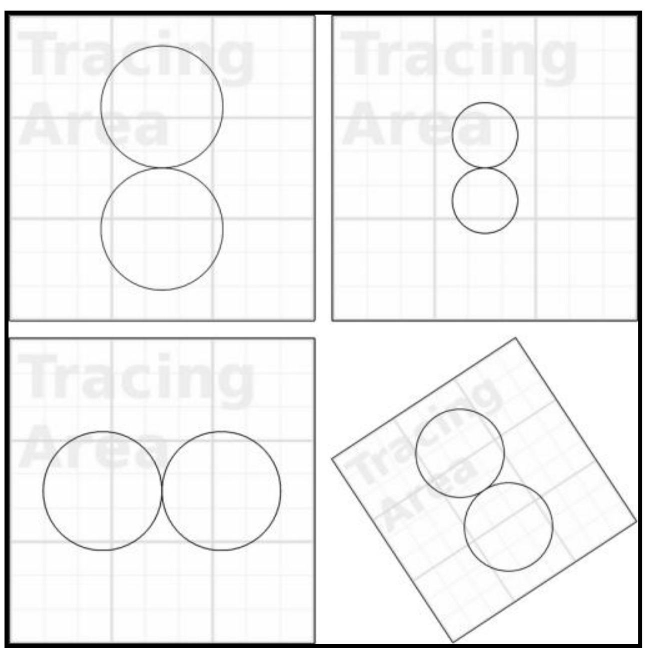


Figure 3 Examples of stimuli. Clockwise from top-left: Baseline, Random Size, Angle Trace, and Infinity.

After completing the drawing tasks, subjects were administered a battery of tests to assess various cognitive aptitudes. These tests included:

(1) *SAT*, subjects self-reported their score on the SAT reasoning test or their ACT score, with ACT scores being converted to an SAT equivalent score for the analysis of results;
(2) *OSPAN or Automated Operation Span Task*, which assessed working memory capacity Unsworth, Heitz, Schrock, & Engle, 2005);
(3) *RAT or Random Associates Task*, which assessed creativity (Mednick, 1963);
(4) *Mental Rotation Task*, which assessed spatial reasoning (Shepard & Metzler, 1971);
(5) *Raven's Progressive Matrices*, which assessed fluid intelligence (Raven, 1958);
(6) *Figure-Comparison Task*, which assessed visual search and motor speed (Salthouse & Mitchell, 1990);
(7) *Einstellung Water-Jug Strategy Task*, which assessed strategy shifting (Tresselt & Leeds, 1953); and
(8) *Shipley's Vocabulary Test*, which assessed verbal abilities (Shipley, 1946).

With the exception of the self-reported SAT/ACT scores, each of the tests was administered by computer.

Table 1 provides a correlation matrix showing the relationship between performance for the cognitive aptitude measures. The most obvious pattern involved four measures (i.e. Vocabulary, RAT, OSPAN and SAT) that tended to each correlate with one another. The Mental Rotation, WaterJug and Ravens tasks did not correlate with the other measures. It was noted that there was insufficient variability between subjects on the WaterJug task for it to be an effective indicator.

Table 1 Correlations for the cognitive aptitude measures.

Pearson Correlation Matrix								
	VOCAB	ROTATION	RAT	SPEED	WATERJUG	OSPAN	RAVENS	SAT
VOCAB	1.000							
ROTATION	0.059	1.000						
RAT	0.386	-0.060	1.000					
SPEED	0.230	-0.125	0.113	1.000				
WATERJUG	0.015	-0.031	0.177	-0.183	1.000			
OSPAN	0.286	0.133	0.364	0.251	-0.099	1.000		
RAVENS	0.286	0.171	0.220	0.200	0.267	0.297	1.000	
SAT	0.568	0.117	0.304	0.275	0.062	0.458	-0.037	1.000

Note: yellow signifies $p<0.01$ and blue $p<0.05$.

The correlations between performance on the cognitive aptitude measures and performance on the different experimental conditions for the line drawing task are shown in Table 2. Of particular note, the SAT and Vocabulary tests both statistically correlated with most of the line drawing conditions. This suggests a generalized aptitude that enabled subjects to effectively cope with the unusual demands placed upon them by the experimental conditions. The OSPAN correlated with performance in three conditions in which it is reasonable to assert that greater working memory capacities could have been beneficial (i.e. No Ink, Random Inter-Stimulus

Interval and Speeded), although interestingly, OSPAN did not correlate with performance for the memory condition. The Figure Comparison Task (i.e. SPEED) correlated with performance in the Angle and Timeout conditions, which both could have benefited from perceptual-motor skills.

Table 2 Correlations between cognitive aptitude measures and performance measures for the line drawing task.

Pearson Correlation Matrix	VOCAB	ROTATION	RAT	SPEED	WATERJUG	OSPAN	RAVENS	SAT
ACC	0.070	-0.126	0.056	0.373	0.030	0.161	-0.282	0.136
ANGLEDRAW	0.316	0.153	0.185	0.192	-0.039	0.235	0.215	0.451
ANGLE	0.302	0.061	0.216	0.254	0.116	0.244	-0.254	0.515
SIZE	0.190	0.047	0.160	0.198	0.105	0.260	-0.323	0.477
INFINITY	0.284	0.053	0.139	0.186	0.101	0.154	-0.393	0.399
MEMORY	0.351	0.054	0.447	0.206	0.007	0.198	0.077	0.426
NO_INK	0.341	-0.029	0.189	0.219	0.173	0.306	-0.091	0.476
PRACTICE	0.120	0.031	0.186	0.191	-0.211	0.187	0.071	0.014
ISI	0.298	0.125	0.114	0.178	0.005	0.365	-0.446	0.456
TIMEOUT	0.305	0.128	0.141	0.251	0.018	0.243	-0.180	0.460
SPEEDED	0.302	0.057	0.241	0.231	0.080	0.314	-0.008	0.443
UNGUIDED	0.102	0.036	0.210	0.103	-0.149	0.169	0.085	-0.021

Note: yellow signifies p<0.01 and blue p<0.05.

In further analyzing these results, the performance in each of the experimental conditions was compared to performance during the initial practice sessions. As shown in Table 3, there was a correlation between how well subjects performed during practice and most of the conditions. However, the notable exceptions were the Angle Draw and Memory conditions. It is believed that these conditions placed demands upon subjects to adapt to unfamiliar task demands and subjects were more or less effective in doing so.

Observation of subject performance revealed three salient strategies that provided the basis for much of the subsequent analysis (see Figure 4). In accordance with these strategies, trials were categorized as either: *Circles*, indicating that the subject drew a circle, after which they lifted their pen and drew a second circle; *Middle*, indicating that they began in the middle and drew a figure-8, *Extreme*, indicating that they began at the top and drew a figure-8; or *Other*, in which their strategy did not correspond to any of the three predominant strategies. Table 4 shows the proportion of trials for which each strategy was observed for each experimental condition.

Table 3 Correlation between performance during initial practice sessions and later performance during experimental conditions.

Pearson Correlation Matrix		
	PRACTICE	UNGUIDED
ACC	0.193	0.395
ANGLEDRAW	-0.013	-0.132
ANGLE	0.286	0.255
SIZE	0.184	0.252
INFINITY	0.250	0.326
MEMORY	-0.033	-0.111
NO_INK	0.208	0.259
ISI	0.318	0.350
TIMEOUT	0.204	0.276
SPEEDED	0.390	0.537

Note: yellow signifies p<0.01 and blue p<0.05.

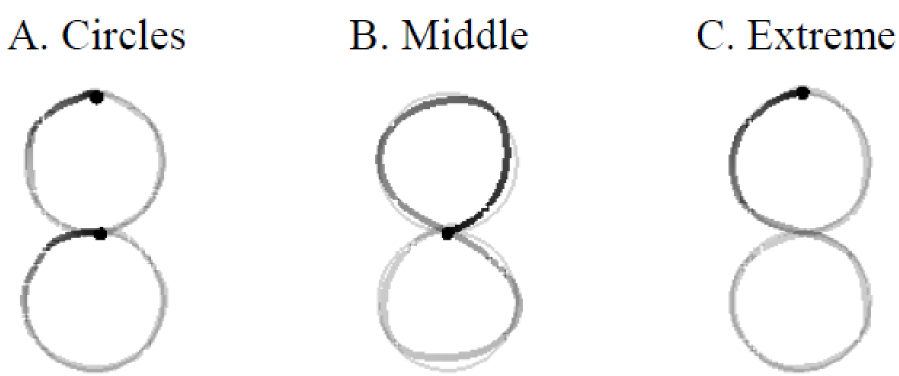


Figure 4 Illustration of the three predominant strategies observed for the Figure-8 Drawing task.

Table 4 Proportion of trials in which strategies were observed for each experimental condition.

Condition	Other	Circles	Middle	Extreme
Practice	0.34	0.24	0.18	0.24
Baseline	0.02	0.26	0.54	0.17
Accuracy	0.00	0.30	0.47	0.23
Angle Draw	0.05	0.58	0.20	0.17
Angle Trace	0.02	0.22	0.70	0.07
Random Size	0.01	0.20	0.70	0.09
Infinity	0.01	0.15	0.75	0.09
Memory	0.17	0.33	0.28	0.21
No Ink	0.02	0.19	0.71	0.08
Random ISI	0.02	0.17	0.61	0.20
Trial Timeout	0.01	0.21	0.57	0.21
Speed	0.02	0.23	0.50	0.25
Total	**0.06**	**0.26**	**0.52**	**0.17**

Variation in strategy selection for the present study was calculated using Shannon's Entropy (Shannon, 1948). Calculations showed that entropy strongly and significantly ($p < .05$) correlated with recurrence ($r = -.989$), repetitiveness of strategy use across trials with higher recurrence indicating greater use of a similar set of strategies, and determinism ($r = -.862$), repetitive patterns in strategy use with high determinism indicating more repetitive patterns.

For each experimental condition, trials were split into blocks representing the initial series of trials and the final series of trials. For example, for a condition with 50 trials, entropy for the initial trials was computed using trials 1-15 and entropy for the final trials was computed using trials 25-50. A 2 × 10 (phase [initial|final] × condition) repeated measures analysis of variance revealed a significant main effect for condition, $F(9, 666) = 80.21$, Mse = .092, $p < .001$, partial $\eta2 = .520$. Bonferroni posthoc tests indicated that there was significantly more entropy in the memory condition than the others ($p < .05$). This was expected, as both the stimuli and task were far more complex than any other condition. Also, the entropy associated with the Angle Draw condition was significantly greater than the Accuracy, Infinity, No Ink, Random ISI, Trial Timeout, and Speed conditions; there were no differences for the other conditions. These findings indicate that for the Memory and Angle Draw conditions, subjects showed a greater tendency to explore alternative strategies for coping with the demands associated with these tasks.

There was also a significant main effect for phase, $F(1, 74) = 68.67$, Mse = .045, $p < .001$, partial $\eta2 = .837$. As expected, initial entropy was significantly higher than final entropy, MINITIAL =

.579, (SE = .017); MFINAL = .366 (SE = .020). With the exception of the memory condition, final entropy was always lower than initial entropy. Table 5 displays the initial and final entropy scores for each condition.

Table 5 Initial versus final entropy.

Condition	**Initial** *M*	*SD*	**Final** *M*	*SD*	**Effect size d**
Accuracy	0.50	0.25	0.20	0.29	1.10
Angle Draw	0.67	0.33	0.42	0.35	0.74
Angle Trace	0.53	0.33	0.32	0.32	0.64
Random Size	0.50	0.30	0.33	0.34	0.54
Infinity	0.51	0.29	0.20	0.33	0.99
Memory	1.09	0.17	1.08	0.18	—
No Ink	0.50	0.28	0.30	0.35	0.62
Random ISI	0.47	0.29	0.26	0.32	0.71
Trial Timeout	0.52	0.29	0.27	0.23	0.99
Speed	0.50	0.26	0.30	0.33	0.71

In general, there was more uncertainty in strategy selection during the initial trials of each condition and less uncertainty during the final trials. Additionally, entropy during the final trials was not zero, indicating that at least some participants were still switching strategies at the end of the block of tests. Finally, and more importantly, with the exception of the Memory and Angle Draw conditions, entropy was largely consistent across conditions. This indicates that it is not the task constraints, but individual differences, that best explain patterns in strategy use.

As shown in Table 6, cognitive aptitude measures were correlated with entropy for initial and final trials for each experimental condition. Verbal ability (Shipley's Vocabulary test and SAT), working memory span (OSPAN), and creativity (RAT) correlated with entropy. The negative correlations between Vocabulary, OSPAN, and SAT indicate that participants with higher verbal ability and executive function demonstrated less variability in strategy selection. Some interesting patterns emerge when one considers initial versus final entropy. Participants with high verbal ability (Vocabulary and SAT) were more likely to settle on a preferred strategy (negative correlation with high entropy). Participants with high working memory span (OSPAN) consistently show lower entropy (negative correlation with both initial and final entropy). There also seems to be a different pattern in creativity and intelligence. The RAT shows a positive correlation with both overall and initial entropy, suggesting a relationship between creativity and initial exploration.

Table 6 Correlation between cognitive aptitude measures and entropy.

ID Measure	Overall Entropy	Initial Entropy	Final Entropy
Vocabulary	-.256**	-.143	-.433***
SAT	-.234*	-.120	-.283**
Ospan	-.419***	-.289*	-.415***
RAT	.205*	.274**	.189
Rotation	-.073	-.041	-.080
Speed	-.093	-.031	-.178
Water Jug	.069	.152	.074
Ravens	.008	.055	-.111

Note *** $p < .01$, ** $p < .05$, * $p < .10$

In summary, differences in creativity, vocabulary, and memory span predicted the variability in strategy use for individual test subjects. Creative participants were more likely to shift strategies and explore different strategies, whereas participants with high verbal ability and memory span were more likely to persevere with their selected strategies.

3. RUMRUNNER MODEL OF STRATEGY SHIFTING

Based on the experimental results reported in the previous section, steps were undertaken to develop a computational model that predicts the likelihood of an individual subject shifting strategies based on task and individual factors. The resulting model combined a number of factors which are summarized in the following sections, and described in detail in Appendix 1.

Task Shift

Task Shift refers to any change in the task that might prompt the adoption of a different strategy. For example, if a football coach loses a star running back due to injury, the coach may adapt by calling more passing plays and fewer running plays. This is probably the most straight-forward aspect of the model as it is clear that when people are faced with new challenges, they are likely to alter their behavior to compensate for next-task demands. For example, when a task becomes more difficult, people are likely to change to a more optimal strategy (Reder, 1987). In RumRunner, the probability of strategy shift increases in proportion to task shift.

Already Doing

Already Doing nullifies the effect of Task Shift in cases where the current strategy is suitable for the task even after the task changes. For example, the coach whose running back is injured might already have been favoring passing plays because the opponents are weaker at defending a pass. Furthermore, people tend to use a particular strategy that has proven effective in the past (Reder & Schunn, 1999). Hence, if a task change biases their previous strategy, it would make sense for them to continue doing it.

Time on Task

The longer a person has been doing a particular task, the greater the probability of switching to a new strategy. Although there is some evidence that under certain circumstances, people will continue with a strategy even after it has become ineffective (e.g. Broder & Schiffer 2006), it is also possible that as familiarity with the task increases, the person may come to a deeper understanding of the task itself, its limits, possibilities, and operation, which allows for the realization that other strategies can be used. RumRunner can accommodate either hypothesis because the coefficient on each model factor may be positive or negative. Time on Task is a task-based factor because at each point in the task, it affects everyone equally and remains agnostic to a person's individual experiences. Here, Time on Task is defined as the number of trials or events relative to the current trial. This is operationalized in RumRunner as a slow-moving power function that uses time on task as its input.

Recently Shifted

If there has recently been a strategy switch, then it is less likely that a person will switch strategies again in the near future. In other words, individuals give each new strategy a trial period to assess its effectiveness. As such, a person will persist with the new strategy, giving it a chance to improve performance. This is a well-known principle that extends back to research on mental set in problem solving in which a person will persist, even when a strategy has become suboptimal (e.g., Luchins, 1942). This principle was implemented in RumRunner as a fast-moving power function with a very strong influence immediately after a strategy shift, which quickly decreases over time. This is in contrast to the Time on Task Factor which has the opposite effect of *increasing* the probability of strategy shift over time, but over a longer timeframe. The net effect is that each strategy is unlikely to be used for either a very short or very long period of time.

Number of Strategy Shifts

It was observed that the more strategy shifts that had occurred, the less likely it was that a person would shift in the future. This may be viewed as a frustration factor because the more often a person has switched strategies in the context of a given task, the more likely they are to be frustrated with their performance, and hence the less likely they are to try a different strategy in the future. We are unaware of any research directly bearing on this idea, but we are proposing it here as a plausible process that can be operating during a given task. This is also implemented as a power function of the number of strategies already tried, with the earlier ones having a greater influence than the later ones. That is, while there is an influence of the number of strategies tried for a task, the fewer strategies that have been tried, the larger the influence that a new strategy will have. However, as the number of different strategy attempts increases, there is a resulting decrease in the cumulative influence of each additional new strategy attributable to an individual becoming increasingly frustrated.

Planning/Preparation

When switching strategies within a task, Luwel (2009) reported a switch cost in the form of longer response times; accuracy, however, was not affected. Thus, increased delay before strategy execution implies impending strategy shift. This is the only factor to use information from the current trial (as opposed to previous trials). In a discrete task, this is operationalized as the amount of planning time prior to producing a behavioral response. In a continuous task, this would be operationalized as a decrease in active performance, consistent with the idea that the person has momentarily disengaged from the task to plan and prepare a new strategy. This factor is a power function of the extra planning time spent on the current trial, relative to the individual's previous trials. This power function increases as planning time becomes further removed from the maximum of the current range of trial times.

Performance Dip

A decrease in performance may motivate a person to shift to a new way of completing the task (Brand, 2008; Lovett & Anderson, 1996; Reder, 1987; Reder & Schunn, 1999). That is, if a person notices that their performance is declining, they will be motivated to change their strategy in order to improve, with the greater the drop, the greater the motivation to change strategies. The performance dip factor is operationalized in the model as a power function, again based on the extent to which the person's performance falls below the minimum level of performance attained in the recent past. The greater the deviation from that minimum, the greater the influence will be.

Bad Shift

The increased likelihood of a strategy shift following a performance dip is even more likely if the dip occurs as a consequence of a strategy shift. We call this factor Bad Shift because the change to a new strategy was unsuccessful. If a new strategy decreases performance, the person will be motivated to revert to the previous one or try yet another. In RumRunner, this factor doubles the probability of an impending strategy shift because it is the result of both a performance dip and a strategy change.

Spasms

In other cases, decreased performance may simply be the result of inattention or random variability. These phenomena are referred to collectively in RumRunner as Spasms. In such cases, people are less likely to switch strategies (as phenomenologically, it will be clear that the performance dip is not due to an ineffective strategy, but rather a lack of attention and effort) and simply reassert their current strategy. We operationalize the presence of a spasm in performance by dividing in half the influence of an increase in the probability of switching strategies that was derived by the performance dip.

No Improvement

Related to the idea that strategy shifts can be caused by dips in performance, it is also possible that a strategy shift can occur if a person has been persisting with a strategy for a period of time or a number of trials without any improvement. When this happens, a plateau has been reached and, under those circumstances, a person may be motivated to try a new strategy to boost performance.

Flexibility

Finally, it is well-known that there are individual differences in strategy selection (e.g., Miller et al., 2002; Rakow et al., 2010; Schaeken, De Vooght, Vandierendonck, & d'Ydewalle, 2000), and so it seems likely that there will be individual differences in the likelihood that a given person will be predisposed to strategy switching. Some people have a more liberal bias, and are more willing to experiment. Alternatively, others are more conservative and are more willing to persist with a strategy, switching only when other factors strongly push them to do so. These biases were illustrated in the previous findings regarding the relationships between certain cognitive aptitudes and the propensity to explore alternative strategies. This is operationalized in RumRunner by providing a running average of the difference between the predicted probability that a person will switch strategies, and whether a strategy shift actually occurred. Note that while the Number of Strategies factor of the model varies with how long a person has been on task, the Flexibility factor is a more stable continuous influence on strategy switching.

3.1 Validation of RumRunner Model

To assess how well the Rum Runner model predicted strategy shifts, it was fit to three data sets, two based on the Figure-8 Drawing Task, and a third from a higher-level decision task.

Figure-8 Drawing Task

As described previously, subjects were given a series of figures to trace (or, in some conditions, draw elsewhere on the screen) using a drawing tablet. The standard figure was two stacked circles (resembling a figure 8), with different variations presented in the two experiments. In one experiment (Curve Drawing 1), the variables were the angle of rotation, the size of the figure, whether the figure was traced or drawn separately, whether a person could see what they were drawing, a horizontal eight that was relabeled as an infinity sign, etc. From this data set, three drawing strategies were identified. These were (a) *circles*, in which a person drew two circles using two strokes, (b) *middle*, in which a person started at the intersection of the circles and drew it in a single stroke from there, and (c) *extreme*, in which a person started anywhere except the intersection of the circles and used a single stroke from there. Together, these strategies accounted for over 98% of all trials.

In a second experiment (Curve Drawing 2), the variables were the angle of rotation, whether the figure was drawn or traced, and three conditions in which the figure-8 was slowly altered over the course of 12 trials to encourage one of three strategies (the morph conditions; see Figure 5). To encourage the circles strategy, the two circles that comprised the figure-8 were slowly separated. To encourage the extreme strategy, the figure was altered so that it was less symmetrical and appeared hand-drawn. Finally, to encourage the middle strategy, the figure was altered such that the lower left part of the figure was disconnected from the intersection and slowly moved away. After data collection, it was apparent that this actually did not encourage people to use a strategy in which they started at the middle crossing point. Instead, people were beginning with the hook end. This was relabeled as the *hook* strategy. Finally, there was a jumble condition in which stimuli from the various prior conditions were randomly presented.

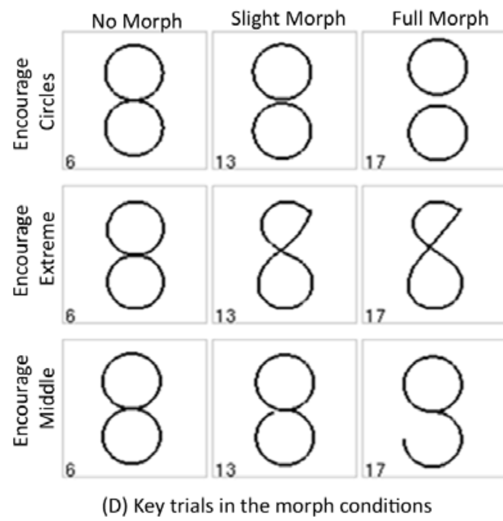


Figure 5 Alterations to figure-8 used to encourage subjects to adopt alternate strategies.

There were 74 participants in the Curve Drawing 1 and 83 in Curve Drawing 2 data set. Participants were undergraduates recruited from the University of Notre Dame and the University of Memphis. The Curve Drawing 1 data set yielded 31,529 trials in 12 conditions, while there were 14,608 trials across eight conditions in the Curve Drawing 2 data set. The base probabilities of trial-level strategy switch were 0.240 and 0.244 in data sets 1 and 2, respectively.

An important difference between the experiments was that the stimulus was largely stable in Curve Drawing 1 (with the exception of minor changes in the angle of the target figure to eliminate practice effects), but greatly varied in Curve Drawing 2 (i.e., the morph and jumble conditions). Consequently, the RumRunner model predicts that there should be a greater influence of experience-based factors in Curve Drawing 1, as compared to Curve Drawing 2.

Binary Choice Task

For this task, subjects were asked to make a series of choices. On each trial, they were asked to select their daily choice for their home temperature setting, how they would listen to music, how they would watch a movie, and which snack they would eat. The choices for the temperature option were 62, 64, 66, 68, 70, 72, 74, 76, 78, and 80 degrees. The choices for the music option were radio, own CD, buy CD, own iPod, download, and live music. The choices for the movie option were TV, Redbox, Blockbuster, download, dollar theater, and theater. Finally, the choices for snack were apple, banana, cookie, chips, and Cold Stone (an ice cream parlor).

At the beginning of the study, subjects were asked to select their ideal preference for each of these three choices. Then, for the primary task, on each trial subjects were to make choices for each of these options to maximize their score. The score was derived using three basic components. These were (a) the most economical choice, (b) a hedonistic ideal, and (c) the time to respond. The most economical choices for each of the options were temperature: 62, music: radio, movie: TV, and snack: chips. To calculate the hedonistic ideal we took the difference between their quality of life scores on the present trial and the scores from their initial preferences; that is there was no preset standard of preference. The default options presented by the task were temperature: 72, music: live music, movie: theater, and snack: Cold Stone. Finally, the performance score was influenced by how long it took a person to complete a trial, with the score decreasing with longer response times.

Participants were 36 individuals who volunteered for monetary compensation on Amazon Mechanical Turk™ (AMT). AMT allows participants to receive small rewards for completing Human Intelligence Tasks (HITs) online. Research has suggested that AMT is a reliable and valid source of experimental data (Paolacci, Chandler, & Ipeirotis, 2010) All participants who completed the study were paid $1.75. The Making Choices data set contained 1404 valid observations with a switch rate of 0.529.

For this task, we identified five strategies. These were (a) *no change* from what was done on the previous trial, (b) change *single* in which subjects changed the value for one choice compared to the previous trial, (c) change *some* in which subjects changed the values for two or three of the choices compared to the previous trial, (d) change all of the values compared to the previous trial, and (e) *speed* in which subjects completed all of the choices at a speed of 8 seconds or less.

RumRunner Model Fitting Procedure

A logistic regression model was used to predict the probability of strategy switching (1 = switch, 0 = no switch) at the trial level from the 11 predictors of the RumRunner model. Due to the nested structure of our data where trials are nested within conditions and conditions are nested within subjects, a mixed-effects logistic modeling approach was adopted (Pinheiro & Bates, 2000). Mixed-effects models include a combination of fixed and random effects. The random effects for the present analyses were *subject*, *condition*, *trial number*, and *number of active components*. Subject was a categorical variable with 74 levels in Curve Drawing 1, 83 levels in Curve Drawing 2, and 36 levels in Making Choices. Condition was also a categorical variable with 12 and 8 levels in Curve Drawing 1 and 2, respectively. Condition was not included in Making Choices because there was only one condition in that Study. Trial number and the number of active components were integers. The number of components was included as a random effect because all 11 components were not applicable in every trial. For example, the Bad Shift component is not applicable when a person does not switch strategies on the prior trial.

The fixed effects were the 11 model factors. Spasm was excluded because transitory spasms were very rare in the data sets. All three task-based factors were included as fixed effects in the Curve Drawing 2 model, but the *Already Doing* factor was excluded from the Curve Drawing 1 and Making Choices models, because we were not explicitly biasing any particular strategy in those studies. The *Task Bias* factor was also excluded from the Making Choices model because the task did not systematically change in that study.

Five models were estimated for each data set, yielding 15 models in all. These included: (a) a *task-based* model with only task-related predictors (i.e. Task Shift, Already Doing and Time on task), (b) an *experience-based* model with only experience-related predictors (Recently Shifted, Number Strategy Shifts, Planning/Preparation, Performance Dip, Bad Shift, Spasm, No Improvement and Flexibility), (c) a *task-experience* model with both task and experience-related predictors, (d) an *intercept-only* model with a fixed intercept, and the random factors, but none of the task or experience-related predictors, and (e) a *random* model, which was identical to the task-experience model, but with a randomly shuffled surrogate of the dependent variable (strategy switch); this model was included as a control

A tolerance analysis was performed to detect potential multicollinearities prior to constructing the logistic regression models. With the exception of Number of Strategy Shifts, Time On Task, and Flexibility (with tolerances in the 0.10 to 0 .27), tolerance values of the remaining eight components exceeded or were very close to the recommended value of 0.4 (Allison, 1999). This indicates that multicollinearity was not a major concern with the predictor set.

RumRunner Model Fitting Results

Receiver operating characteristic (ROC) curves obtained from the 15 models are shown in Figure 6. For Curve Drawing 1, the random model yielded the poorest fit, while the experience-based and the task-experience models yielded the best fits. Fits obtained by the task-based and intercept-only models were between these two extremes. Thus, the task-based predictors were minimally relevant, and it is the experience-derived factors that best predicted strategy switching in the Curve Drawing 1 data set.

The model-fitting procedure yielded a somewhat similar outcome for Curve Drawing 2. As expected, the ROC curves indicate that the random model yielded performance that was lower than that expected by chance (chance = .5 for ROC curves). Similar to Curve Drawing 1, the experience-based and task-experience models yielded the best performance, while performance was lower for the task-based and intercept-only models. One difference in Curve Drawing 2 was that the task-based model outperformed the intercept-only model. Although the difference in AOC values was small, a likelihood-ratio test indicated that the difference in log likelihoods associated with these models was significant $\chi^2(2) = 286.8$, $p < .0001$. Another difference was that the task-experience model yielded a better fit than the experience-based model in Curve Drawing 2, $\chi^2(3) = 676.8$, $p < .0001$.

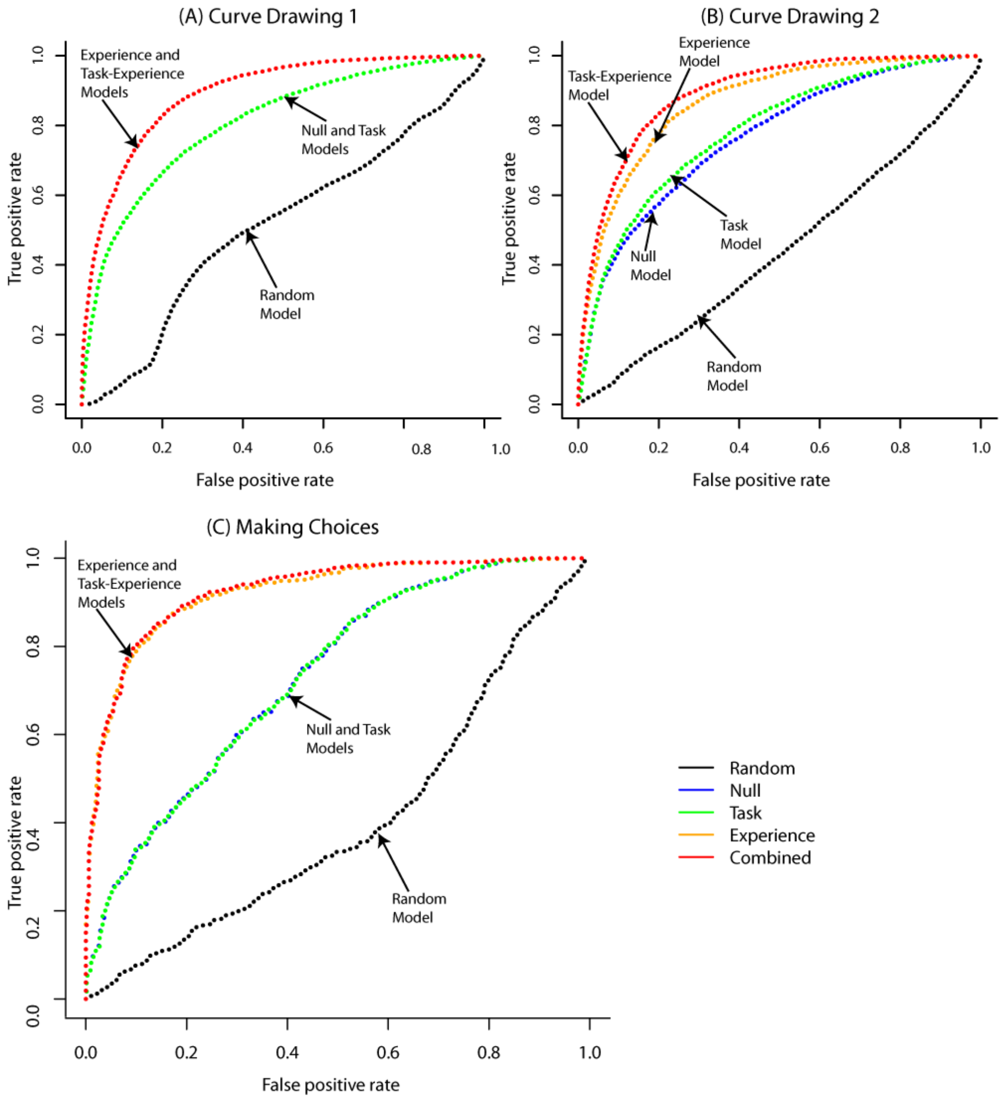


Figure 6 Receiver Operator Curves for RumRunner modeling fitting.

The pattern of results for the Making Choices data set was very similar to the Curve Drawing 1 data set. Specifically, the random model yielded the poorest fit, the experience-based and task-experience models had the best fits, and fits associated with the task-based and null models were in between these two extremes. Taken together, the results from all three data sets indicate that the experience-derived factors were the most diagnostic predictors of strategy switching. Although the task-based predictors yielded negligible effects in Curve Drawing 1 and Making Choices, they had some impact in Curve Drawing 2, where the task was more explicitly and systematically manipulated.

RumRunner Classification Accuracy

The subsequent discussion focuses on the task-experience models for all data sets, as these models encompass all the predictors and outperformed the random and intercept-only models. Classification tables were obtained by comparing the predictions generated by these models to observed strategy shifts. An analysis of the ROC curves indicated that a probability threshold of 0.6 was optimal in separating trials where a strategy shift occurred (probability > 0.6) from trials where there was no shift in strategies (probability <= 0.6). The 0.6 cutoff is optimal because it maximizes the true-positive rate while minimizing the false-positive rate.

A number of conclusions can be drawn from the classification tables (See Table 7). First, the models accurately predicted strategy switches in 66% of the cases in both curve drawing experiments; this exceeds the 24% base-rate of strategy shifting. Accuracy of predicting strategy switches was substantially higher (91.1%) for the Making Choices study (base rate = 53%). The models also accurately predict when a participant uses the same strategy with an accuracy rate of 90% in both curve drawing experiments. Accuracy of predicting no strategy switch was 78% for Making Choices. Taken together, classification accuracy (computed from the diagonals of the classification tables) was 84.3%, 84.1%, and 84.9% for Curve Drawing 1, 2, and Making Choices data sets, respectively.

Table 7 Classification tables for task-experience models.

	Curve Drawing 1			**Curve Drawing 2**			**Making Choices**		
		Pred.			*Pred.*			*Pred.*	
	Obs.	0	1	*Obs.*	0	1	*Obs.*	0	1
N	0	21600	2368	0	9943	1100	0	515	146
	1	2573	4988	1	1221	2344	1	66	677
%	0	90.1	9.90	0	90.0	10.0	0	77.9	22.1
	1	34.0	66.0	1	34.2	65.8	1	8.88	91.1

Although these results suggest that Rum Runner is quite effective in predicting strategy switching, one problem with the analyses was that the models were constructed and validated on the entire data set. This limits claims of generalizability of the models to new trials from new individuals. We addressed this concern by assessing the classification accuracy of the three task-experience models with a between-subjects split-half evaluation method. The analyses proceeded as follows for each data set. Half of the participants were randomly selected and their data points were assigned to the training set. Data from the remaining participants were assigned to the test set. A logistic regression model was constructed from the training data and was used to generate predictions on the testing data. Classification accuracy associated with this split-half evaluation procedure was 82.6% for both curve drawing experiments. This is comparable to the 84% accuracy obtained from the training set alone.

There was a reduction in split-half classification accuracy for the Making Choices data. Here, split-half classification accuracy was 74.4%, which is somewhat lower than the 84.9% accuracy obtained from the entire training set. We suspect that this reduction in accuracy is a consequence of the smaller number of data points in this data set (approximately 1,500 data points in Making Choices compared to the approximately 14,500 and 31,500 data point in Curve Drawing 1 and 2, respectively). Hence, split-subjects accuracy is expected to increase with more subjects and data points. Despite this qualification, it is important to note that this 74.4% split-half accuracy is greater than what could be expected by chance (52.0% base rate). Taken together, these results indicate that RumRunner is indeed a moderately accurate model of strategy switching that is robust across differences in the three studies and generalizes to new individuals.

RumRunner Model Parameters

Table 8 lists the fixed-effects estimates of the task-experience models for all three data sets. The RumRunner components were standardized prior to computing these models in order to afford comparisons across parameters that vary in scale. Parameters with *B* weights greater than 0.5 are emphasized in bold. We note that *Number of Strategy Shifts*, *Flexibility*, and *Time on Task* are influential positive predictors of strategy switching in all three data sets. The direction of the latter two components is consistent with our predictions. Specifically, people with an intrinsic bias to switch strategies (*Flexibility*) are more likely to switch and the probability of switching to a new strategy increases as a function of how long a person is engaged in the task (*Time on Task*). In contrast, we predicted that a person will be less likely to switch in the future if considerable strategy switching is currently occurring. This predictions was not supported because *No. Strategy Shifts* was a positive instead of a negative predictor.

Table 8 Fixed-effects parameter estimates for task-experience models.

	Curve Drawing 1		**Curve Drawing 2**		**Making Choices**	
Fixed Effect	*B*	*SE*	*B*	*SE*	*B*	*SE*
Intercept	-1.95	.399	-1.62	.294	.600	.476
Experience-based						
Z Recently Shifted	-.109	.017	-.080	.027	**.603**	.104
Z No. Strategy Shifts	**.932**	.046	**1.63**	.072	**6.12**	.372
Z Planning/Preparation	.194	.019	.197	.026	ns.066	.134
Z Performance Dip	ns.044	.030	ns.042	.040	ns.041	.149
Z Bad Shift	.209	.024	.134	.032	ns -.033	.154
Z No Improvement	ns.011	.024	.122	.031	ns.106	.102
Z Flexibility	**2.23**	.036	**1.91**	.043	**6.84**	.352
Task-based						
Z Task Biases	.515	.066	.645	.042	-	-
Z Already Doing	-	-	**1.00**	.049	-	-
Z Time on Task	**.680**	.223	**1.70**	.102	**1.78**	.305

Notes. All parameters significant at $p < .05$ unless noted by ^{ns}. Fixed-effects were standardized prior to constructing logistic models

In contrast to these three components that were highly predictive in all three data sets, the predictive power of three other components was tied to one or two data sets. These include *Recently Shifted,* which was dominant in the Making Choices data set, and *Task Biases* and *Already Doing* which were predictive when task constraints changed as was the case in the Curve Drawing data sets. Our prediction that people will be less likely to shift strategies when they have recently adopted a new strategy was also not confirmed by the fact that *Recently Shifted* was a positive instead of negative predictor of strategy switching. In contrast, our prediction that task biases would induce strategy switching was supported because *Task Bias* positively predicted switching.

In addition to the aforementioned six components that were consistently or contextually predictive of strategy switching, four components yielded either non-significant or significant but small effects across all three data sets. These include *Planning/Preparation*, *Performance Dip*, *Bad Shift*, and *No Improvement*. It is possible that these components might play a more substantial role in alternate data sets, a possibility that warrants testing the model on additional data sets.

Finally, it is important to note that our categorization of the components as *consistently*, *contextually*, and *minimally* predictive does not mean to imply that one or two predictors are driving the overall predictions. In contrast, the models are quite robust to small parameterization changes. This was confirmed with a simple sensitivity analysis where model fits were assessed after components were individually removed. With the exception of flexibility in the Making Choices data set, individual component removal had negligible impacts on the model (AOC values always exceeded .8), so there is some confidence that a single component does not bias model performance.

4. STRATEGY SWITCHING IN MULTI-TASKING

Although considerable research has focused on the individual differences that predict multitasking ability, relatively little research has been devoted to multitasking adaptability. We operationalize multitasking ability as some metric of performance when the difficulty level of the individual tasks are at some baseline. In contrast, adaptability refers to a person’s capacity to adapt to changing task constraints. That is, how is performance impacted if the difficulty of one or more of the tasks increases or if a new task is introduced. Therefore, ability and adaptability, although related, are not necessarily the same construct.

Branscome and Grynovicki (2007), assessed adaptability using SynWork (Elsmore, 1994), a multitasking environment with four tasks (memory, math, auditory, and monitoring). They also added a military target-identification task that involved identifying friendly vs. enemy targets. The three (counterbalanced) conditions in their study consisted of: all four SynWork tasks, three SynWork tasks plus the target-identification task, and all four SynWork tasks plus the target-identification task. They found that the third (most difficult) condition was associated with a significant drop in performance compared to the other two conditions. Additionally, Wang, Proctor, and Pick (2007) found that when certain SynWork tasks were biased via a payout structure (i.e., more points were awarded for some tasks), some people were able to strategically adapt to the new task constraints. This suggests that some individuals show a greater propensity for multitasking adaptability than others. Identifying the individual differences associated with multitasking adaptability is the primary goal of this paper.

Another goal of this study was to identify groups of individuals on the basis of their adaptability profiles. A framework to conceptualize adaptivity may be useful in order to distinguish between changes in performance (from a baseline) under increased task difficulty (Target task) versus constant task difficulty (Off-Target task). Such a framework is presented in Figure 7. The vertical axis represents performance on a difficult task(s) relative to a given baseline, whereas the horizontal axis represents performance on a task(s) with baseline difficulty. The different regions in the figure represent various adaptability profiles. It is important to note that “Target” and “Off-Target” do not necessarily refer to the importance of the task, but only to the difficulty level relative to a baseline. For example, if a person is driving while talking on the phone and they receive another call, the difficulty of the phone task has increased so it is the Target task, and driving is the Off-Target task (even though it remains the primary task in terms of importance).

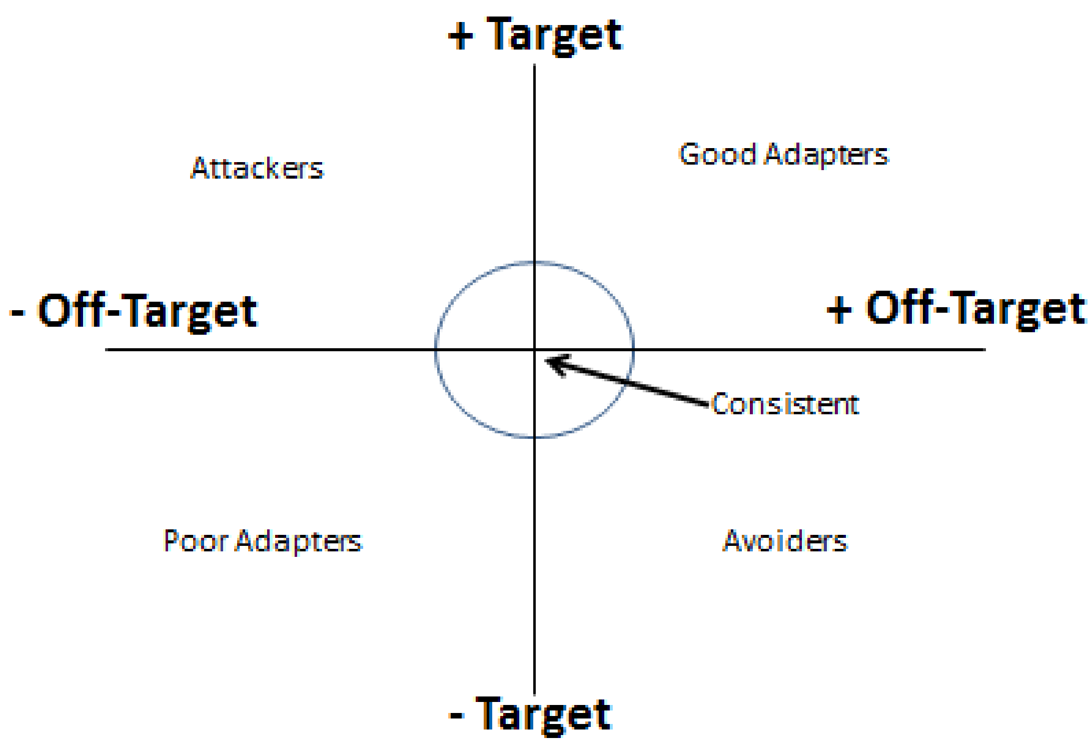


Figure 7 Framework to characterize multi-tasking adaptability.

Five major adaptability profiles are noted in Figure 7. Individuals whose performance is not compromised when additional difficulty in a task(s) is encountered are consistent performers (circle around the origin in Figure 1). The top-right quadrant encompasses the individuals whose performance increases in both the Target and Off-Target tasks (good adapters). Conversely, the bottom-left quadrant represents individuals who show a decline in performance in both Target and Off-Target tasks (poor adapters). There might also be situations when an individual sacrifices performance in one task(s) for the sake of another. When encountering additional difficulty on a task(s), some individuals may choose to tackle the difficult task(s) at the expense of the other task(s). These individuals are referred to as attackers (top-left quadrant). Alternatively, others may neglect the difficult task(s) and focus on the other tasks which are at baseline difficulty. These are the avoiders (bottom-right quadrant).

There are two important points to note about the proposed adaptability framework. First, the distribution of individuals to adaptability profiles need not be uniform. For example, there may be no avoiders for a given multitasking context, whereas all individuals might be classified as avoiders for a different context. In the latter case, the differences among the individuals may still be useful, because all avoiders are not equivalent, and the relative differences among avoiders can be quite informative.

Second, the five adaptability profiles listed above are sensitive to differences in multitasking contexts. That is, a poor adapter in context A is not necessarily a poor adapter in context B. Quite different from a rigid dispositional characterization of an individual's ability to adapt, the primary purpose of the framework is to organize the relationship between individuals or groups within a given multitasking context.

The present study addressed the relationship between cognitive faculties and multitasking ability and adaptability in a pilot simulation task. We collected scores on standard cognitive measures and correlated these with performance under baseline difficulty (ability) and the individual change in performance when task(s) difficulty increased (adaptability). We also grouped individuals based on multitasking adaptability and investigated whether these groups could be differentiated on the basis of the cognitive measures associated with ability and adaptability.

4.1 Methodology for Multi-Tasking Study

Subjects consisted of 32 participants either enrolled in a Midwestern university or volunteers of various educational backgrounds from a Southern city in the United States. At the start of the experiment, participants completed a computer-administered battery of cognitive aptitude measures assessing scholastic aptitude, working memory, creativity, and spatial ability. Specific measures included: SAT or ACT score, OSPAN, RAT, and Mental Rotation Task.

Multi-tasking performance was assessed using the NASA Multi-Attribute Test Battery or MAT-B (Comstock & Arnegard, 1992). The MAT-B is a computerized flight simulator that requires users to simultaneously attend to four individual tasks: System Monitoring, Communications, Resource Management, and Tracking. Each individual task had four difficulty levels: automatic (0), low (1), medium (2), and high (3). The MATB interface is shown in Figure 8 and the following sections describe individual components of the task. Performance was calculated as a product of the scores on the four individual tasks (the composite score) and was displayed to participants via a performance gauge at the bottom of the screen. Therefore, completely neglecting any one task would yield a composite score of zero. Performance scores can range from 0 to 100.

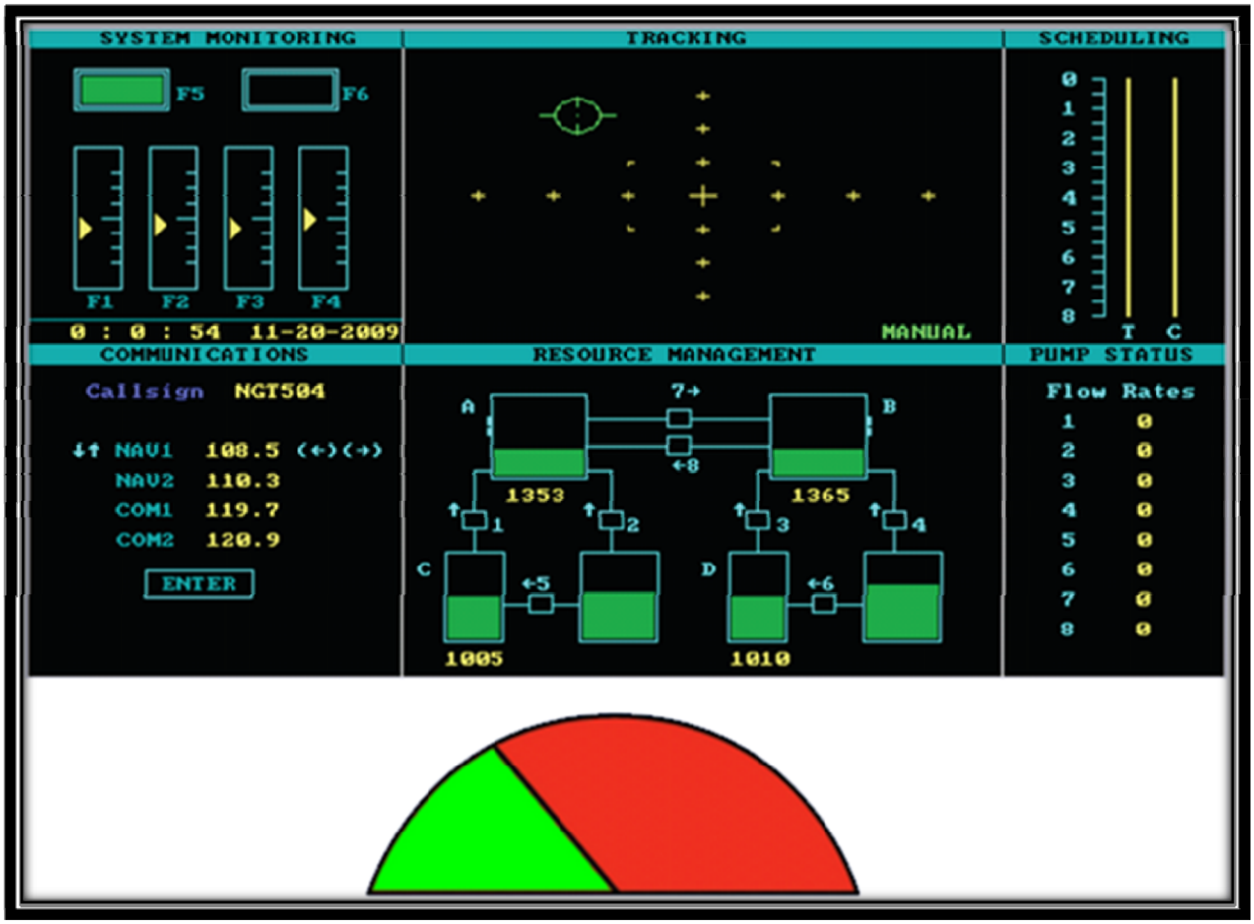


Figure 8 The MATB interface, with the composite score displayed at the bottom.

System Monitoring

In the top-left quadrant of the screen, participants were asked to respond to feedback from lights and gauges. There were two lights at the top of the quadrant: a green light and a red light. Participants were instructed to press the F5 key if the green light turned off and to press the F6 key if the red light came on. Doing so turned the green light on and turned the red light off, respectively.

Beneath the two lights were four gauges, each associated with a corresponding key on the keyboard. Each gauge also had a yellow pointer that typically hovered around the center line. Participants were asked to press the corresponding key (F1 to F4) if any gauge's pointer exceeded one unit in either direction of the gauge's center line. When a correct action for a gauge was performed, participants would receive feedback via a yellow line that would briefly appear on the respective gauge.

Communications

In the bottom-left quadrant, participants were given an identifying call sign (NGT504) and asked to follow audio instructions directed to their call sign while ignoring instructions for other call signs. Each message began with a six-character call sign (which was repeated), followed by a command to change one of four ports (two navigation and two communication) to a particular frequency, represented by a four-digit number. An example command would be, "NGT504, NGT504, change navigation two to one-one-zero-point-five." Participants selected between the four navigation and communication ports with the up and down arrow keys, and modified the frequency using the left and right arrow keys.

Resource Management

In the bottom-right quadrant, participants were asked to manage the fuel levels (represented in green) of two tanks, A and B. The fuel levels in tanks A and B decreased constantly as the fuel was used, but participants were instructed to keep the fuel level of both tanks between the tick marks indicated on each tank (5/8ths full). This was done by turning various pumps on or off to transfer fuel from another tank. Tanks C and D had a finite supply, whereas the other two tanks had an unlimited capacity. To turn a pump on or off, participants pressed the number key corresponding to the pump flow indicators on the screen (one through eight). At various points, one or more pumps would 'fail' and become unusable for a set period of time.

Tracking

In the top-right quadrant, participants were asked to keep a moving reticle (crosshairs) as close as possible to the center crosshairs using a standard joystick. The reticle would drift and randomly change directions with varying force.

Scheduling and Pump Status.

The Scheduling and Pump Status zones on the far right provide supplementary information for the Communications and Resource Management tasks, respectively. They were not necessary for completing any of the tasks, so they will not be discussed here.

4.2 MAT-B Difficulty Levels

Participants first completed the cognitive aptitude measures on one computer for approximately one hour. After a short break, they completed the MATB task on a separate computer for another hour. Participants were seated and used a keyboard and mouse (cognitive aptitude measures) or joystick (MATB), and used speakers or headphones for audio output. The MATB had five separate conditions: Practice, Baseline, Single Difficulty, Paired Difficulty, and Difficulty Ramp-Up. Participants were not given instructions regarding priority of the individual tasks nor were they told how the composite score was calculated.

Practice

The first condition consisted of four separate practice sessions wherein the participants received instructions from the experimenter and then completed each task individually for a total of nine minutes. The composite score was displayed at the bottom of the screen in all practice conditions. Subsequent conditions had participants attend to all four tasks simultaneously, Baseline (BL). The Baseline condition had all four tasks set at the low difficulty level for five minutes. Participants were not given feedback on their scores to encourage them to concentrate on completing all four tasks simultaneously. Subsequent conditions provided feedback on the composite score.

Single Difficulty (SD)

The Single Difficulty condition began with a three-minute warm-up during which all tasks were at baseline difficulty, after which the difficulty level of one task was set to hard (Target task) for one minute, while the others were at easy difficulty (Off-Target). All difficulties were then set to easy for another minute. This was done for all four tasks in succession, then the process was repeated, for a total of 16 minutes.

Paired Difficulty (PD)

The Paired Difficulty condition raised the difficulty of both the System Monitoring and Communications tasks to hard for two minutes (Target), while the Resource Management and Tracking tasks remained at easy difficulty (Off-Target). System Monitoring and Communications were selected as Target tasks in this condition because both provide discrete events which immediately impact participants' scores, thereby forcing them to constantly attend to these tasks.

Difficulty Ramp-Up (Ramp-Up, RU)

For the final condition, after a one-minute warm-up, the difficulty of all four tasks were raised to medium for one minute (RU2), and to the hardest difficulty for one additional minute (RU3).

4.3 Multi-Tasking Ability

Table 9 displays the mean scores and standard deviations for each task in each condition, as well as an average across tasks. There did not appear to be major differences in the average scores across conditions, indicating that participants were able to adapt to increased difficulty.

In order to assess which of the cognitive aptitude measures predicted performance on the MATB, we correlated the scores of each measure with average performance in each condition (see Table 10). Consistent with previous research, the results indicated that scholastic aptitude and working memory were associated with superior performance on the MATB task (ability).

Table 9 Means and standard deviations of MATB scores for each task and condition.

	Task				
Condition	**Monitoring**	**Comms.**	**Resource**	**Tracking**	**Average**
Baseline	93.0 (3.35)	85.0 (16.6)	82.1 (12.3)	84.3 (5.35)	86.1 (9.4)
Single Difficulty					
Monitoring	**67.6 (9.19)**	81.9 (20.9)	82.2 (18.5)	83.4 (8.78)	78.8 (14.3)
Communications	94.9 (4.77)	**84.1 (17.7)**	81.4 (22.8)	85.8 (5.21)	86.6 (12.6)
Resource	93.7 (4.05)	81.6 (20.2)	**80.3 (20.6)**	86.7 (3.95)	85.6 (12.2)
Tracking	94.5 (6.25)	78.9 (23.8)	84.0 (21.0)	**75.9 (7.18)**	83.3 (14.6)
Paired Difficulty	**86.0 (7.67)**	**86.9 (18.8)**	85.7 (11.6)	85.5 (4.85)	86.0 (10.7)
Ramp-Up 2	**91.7 (5.33)**	**91.8 (12.9)**	**89.8 (11.9)**	**78.1 (8.08)**	87.8 (9.6)
Ramp-Up 3	**86.3 (9.89)**	**86.0 (18.7)**	**86.7 (16.9)**	**70.5 (8.40)**	82.4 (13.5)

Note. Bolded values represent tasks with increased difficulty (i.e., the Targets)

Table 10 Correlations between cognitive aptitude measures and MATB scores in each condition.

Measure	**Baseline**	**Single Difficulty**	**Paired Difficulty**	**Ramp Up 2**	**Ramp Up 3**
Aptitude	****.436**	****.447**	*****.503**	***.381**	***.337**
Spatial Ability	-.052	*.343	.247	.308	**.402
Creativity	-.169	*-.364	-.146	.023	-.005
Working Memory	***.353**	*****.648**	***.394**	**.203**	***.356**

Notes. * $p < .10$ ** $p < .05$ *** $p < .01$

4.4 Multi-Tasking Adaptability

The correlations presented in Table 10 represent the cognitive aptitudes associated with multitasking ability. However, it remains to be seen if these are the same abilities that govern adaptability. To answer this question, partial correlations on cognitive aptitude measures and

MATB scores were performed using Baseline performance as a covariate (Table 11). These correlations indicate which cognitive faculties predict performance when task difficulty increases after accounting for general multitasking ability (Baseline scores).

The partial correlations presented in Table 11 were illuminating in a number of respects. First, scholastic aptitude was not significantly predictive of adaptability (change in performance under increased task difficulty). With the exception of the Single Difficulty condition, working memory also played no role in predicting adaptability. Interestingly, spatial ability was a significant predictor of adaptability. Taken together, this pattern indicates that the cognitive aptitudes that predicted general performance (ability) were distinct from those which predicted adaptability. The next step was to determine if we could identify different types of adapters both within and across experimental conditions.

Table 11 Partial correlations between cognitive aptitude measures and MATB scores after controlling for Baseline scores.

Measure	Single Difficulty	Paired Difficulty	Ramp Up 2	Ramp Up 3
Aptitude	.224	.324	.167	.113
Spatial Ability	*****.514**	***.399**	****.461**	*****.563**
Creativity	-.253	.024	.160	.124
Working Memory	***.599	.095	-.070	.083

Notes. * $p < .10$ ** $p < .05$ *** $p < .01$

Single Difficulty (SD)

Difficulty of each of the four tasks in the experimental conditions was either increased from the Baseline difficulty (Target) or equal to the Baseline difficulty (Off-Target). We calculated delta scores for each of the four Target tasks in the SD condition by subtracting the corresponding Baseline score for that task. The four delta scores were then averaged together to yield an average delta Target score. For example, according to Table 10, the average delta Target score for Single Difficulty would be: ((67.6 - 93.0) + (84.1 - 85.0) + (80.3 - 82.1) + (75.9 - 84.3)) / 4 = -9.13 . The average delta Off-Target score would be computed from the 12 Off-Target scores (non-bolded Single Difficulty scores in Table 10) in a similar fashion. Of course, the actual computations of Target and Off-Target delta scores were performed at the subject level, instead of the aggregate scores as in the example above. In this fashion, there was one average delta Target and one average delta Off-Target score for each participant in the Single Difficulty Condition.

After computing the Target and Off-Target delta scores, we used a 2-dimensional k-means cluster analysis to identify different types of adapters in this study. The results of the clustering are presented in Figure 9. We used a k of 3 because preliminary testing indicated that a 3-cluster solution yielded the best separability of the data.

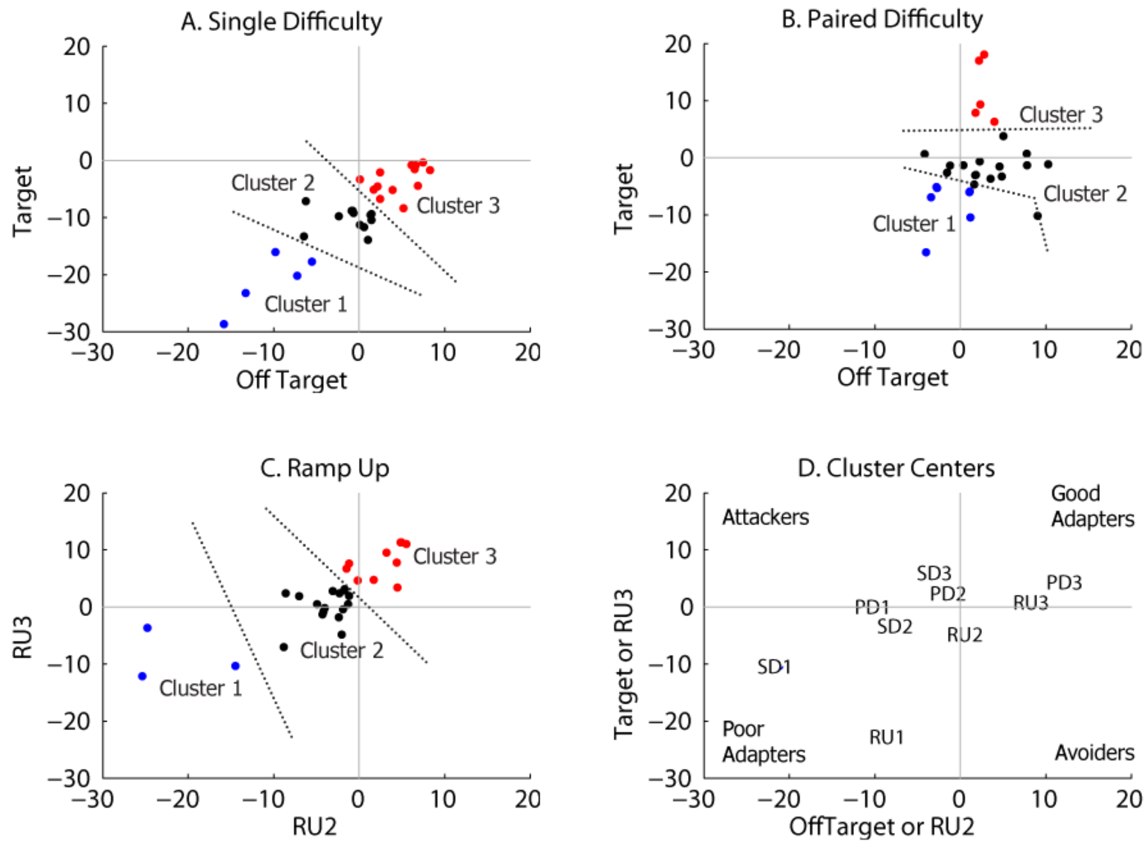


Figure 9 Cluster analyses for all three conditions.

We tentatively identified these clusters as Clusters 1, 2, and 3. In order to differentiate the clusters in terms of their performance on the Target task, a one-way between-subjects ANOVA on delta Target scores with Cluster as a three-level independent variable was performed. The model was significant, $F(2, 27) = 72.5$, $p < .001$, and Bonferroni post-hoc tests revealed the following pattern in the data: Cluster 3 > Cluster 2 > Cluster 1 (See Table 12). Similarly, an ANOVA on the delta Off-Target scores was also significant, $F(2, 27) = 46.6$, $p < .001$. Bonferroni post-hoc tests indicated that Cluster 3 > Cluster 2 > Cluster 1. In line with these findings, we refer to Clusters 1, 2, and 3, as the Low, Medium, and High adaptability groups, respectively.

It is important to note that the clusters in the Single Difficulty condition did not differ on their Baseline scores, $F(2, 27) = 1.20$, $p = .316$. That is, each cluster was equivalent in terms of baseline multi-tasking ability, but there were differences in terms of adaptability.

Table 12 Means and standard deviations (SD) for delta scores in each cluster.

Condition	N	Delta Target	Delta Off-Target	Delta RU-2	Delta RU-3
Single Difficulty					
C1 (Low)	13	-21.2 (4.98)	-10.3 (4.22)		
C2 (Mid)	12	-10.3 (1.96)	-0.97 (2.77)		
C3 (High)	5	-3.47 (2.49)	4.60 (2.61)		
Paired Difficulty					
C1 (Low)	7	-8.04 (4.15)	-1.38 (2.37)		
C2 (Mid)	16	-2.03 (2.97)	3.36 (4.06)		
C3 (High)	5	11.7 (5.44)	2.61 (0.842)		
Ramp-Up					
C1 (Low)	3			-14.4 (7.83)	-32.2 (6.51)
C2 (Mid)	16			.186 (4.92)	-6.27 (4.24)
C3 (High)	10			13.2 (5.12)	4.55 (4.23)

Note. Number of participants are not equal across conditions because occasional outliers were removed. C1, C2, C3 = Cluster 1, 2, and 3, respectively.

Paired Difficulty (PD)

The analyses for the Paired Difficulty condition proceeded in a similar fashion. For PD, the System Monitoring and Communications tasks were the Target tasks, whereas the Resource Management and Tracking tasks were Off-Target. The clustering yielded 3 clusters as illustrated in Figure 9.

The ANOVA on delta Target scores was significant, $F(2, 25) = 41.4$, p <0 .001. Bonferroni post-hoc tests revealed the following pattern in the data: Cluster 3 > Cluster 2 > Cluster 1. The ANOVA on the delta Off-Target scores was also significant, $F(2, 25) = 4.90$, p =0 .016. Bonferroni post-hoc tests indicated that Cluster 2 had greater delta Off-Target scores than Cluster 1. Cluster 3 was not statistically significantly different from Clusters 1 or 2 (see Table 12). As with the SD clusters, the PD clusters did not differ on their Baseline performance, $F(2, 25) = 0.458$, p =0 .638, indicating that the clusters are distinguishing individuals on the basis of adaptability rather than general ability.

Difficulty Ramp-Up.

For the Difficulty Ramp-Up condition, all four tasks in Ramp-Up 2 and Ramp-Up 3 were raised in difficulty, thereby negating the Target and Off-Target distinction. Our delta scores were calculated by subtracting the Baseline scores from both the RU2 and RU3 conditions and clusters were derived in same fashion described above. Therefore, these delta scores represent performance in the face of moderate (delta RU2) and high (delta RU3) difficulty.

The ANOVA for the RU2 delta score was significant, $F(2, 26) = 39.88$, $p <0 .001$, and Bonferroni post-hoc tests revealed the following pattern in the data: Cluster 3 > Cluster 2 > Cluster 1 (see Table 12). The ANOVA for the RU3 delta score was also significant, $F(2, 26) = 74.98$, $p <0 .001$, and Bonferroni post-hoc tests again indicated that Cluster 3 > Cluster 2 > Cluster 1.

Unlike the SD and PD conditions, there was a difference in baseline performance among the Clusters in the Difficulty Ramp-Up condition $F(2, 26) = 6.86$, $p =0 .004$), where Cluster 2 (M = 78.5) outperformed Cluster 3 (M = 66.0). However, there was no difference between the clusters for the warm-up phase of the Ramp-Up condition (RU1), $F(2, 26) = 0.841$, $p =0 .443$. This indicates that the clusters exhibited equal performance before the difficulty increased. Our analyses used the Baseline score as the reference in order to stay consistent with the SD and PD conditions.

Clustering Consistency Across Conditions

One remaining question is if the individual membership in each cluster is consistent across conditions. For example, does the High adaptability cluster in one condition contain the same individuals as the High adaptability cluster in another condition? Failure to find consistencies across conditions might complicate (though not invalidate) any further analyses on the clusters. Fortunately, Spearman correlations on cluster membership (see Table 13) indicated that cluster membership was, in fact, consistent across conditions. 37% of the individuals were in the same cluster (High, Medium, or Low) across all three conditions, while 63% of the individuals were in the same cluster in two of the three conditions. No individual was assigned to different clusters in all three conditions. This suggests that there was some consistency in cluster membership across conditions.

Table 13 Spearman correlations between clusters and conditions.

	Paired Difficulty	Difficulty Ramp-Up
Single Difficulty	.621***	.395**
Paired Difficulty		.447**

Notes. * $p < .10$ ** $p < .05$ *** $p < .01$

Predicting Adaptability Type

After grouping individuals into the different clusters, our next analysis was to determine whether the three clusters in each condition differed on the basis of any cognitive measures. Based on the partial correlations (see Table 11), we would expect that spatial ability would distinguish the clusters, and that the other cognitive aptitude measures would have no impact. We performed separate between-subjects ANOVAs for each condition with Cluster as the independent variable with 3 levels (for High, Medium, and Low) and spatial ability scores as the dependent variable. The analyses yielded significant models for Single Difficulty, $F(2, 22) = 4.89$, $p =0\ .018$ and Paired Difficulty $F(2, 22) = 4.02$, $p =0\ .032$, and a marginally significant model for Difficulty Ramp-Up $F(2, 21) = 2.77$, $p =0\ .086$.

Descriptive statistics for the spatial ability measure are displayed in Table 14. Bonferroni post-hoc tests indicated that individuals assigned to the High adaptability cluster in the SD condition had higher spatial ability scores than individuals assigned to the Medium adaptability cluster (High > Medium). Individuals in the High adaptability cluster also had higher spatial ability than individuals in the Low adaptability cluster for the PD condition (High > Low). This High > Low pattern was replicated in the RU condition but was only marginally significant ($p =0\ .090$). The ANOVAs were repeated with the other four cognitive measures (scholastic aptitude, working memory, and creativity) as dependent variables. Importantly, none of the models were statistically significant in any condition ($p >0.05$), further indicating that these measures did not play a major role in discriminating individuals on the basis of adaptability.

Table 14 Spatial ability means (standard deviation in parenthesis) for clusters in each condition.

Condition	Low	Medium	High
Single Difficulty	3.50 (4.51)	2.60 (5.95)	10.5 (6.43)
Paired Difficulty	3.60 (4.83)	6.40 (5.79)	12.8 (4.15)
Difficulty Ramp-up	-1.50 (2.12)	7.64 (5.72)	8.88 (5.84)

5. CONCLUSION

The Robust Automated Knowledge Capture LDRD was successful in that the project advanced a scientific understanding of individual differences in cognitive performance. The primary product of the LDRD has been the RumRunner model which identifies the task and experience related factors that predict an individual's propensity for strategy shifting. This is important because many national security applications place individuals in situations where they must appraise complex, often ambiguous, conditions and decide whether they should persist with their current strategy or abandon their current strategy in favor of an alternative strategy. RumRunner provides the basis for assessing performance on a decision-by-decision basis to ascertain how various factors are likely to impact subsequent decision making. Furthermore, by applying RumRunner as a basis for evaluating expert decision making, it may provide insight into the factors that distinguish expert from novice decision makers within a given domain.

Based on the research conducted through the Robust Automated Knowledge Capture LDRD, individual measures of cognitive performance have been identified that are believed to provide a basis for assessing individual adaptability. Here, adaptability refers to an individual's capacity to effectively recognize and cope with changing circumstances. In particular, three measures have been identified: (1) sensitivity to negative feedback in association with changing stimulus conditions; (2) sensitivity to changing contingencies underlying positive feedback and (3) mental flexibility as measured through the Mental Rotation Task. Subsequent work building on the success of the Robust Automated Knowledge Capture LDRD will focus on incorporating these measures into a test battery to provide a tool for assessing individual capacity for adaptive thinking.

6. References

Allison, P. D. (1999). Multiple Regression. Thousand Oaks, CA: Pine Forge Press.

Brand, M. (2008). Does the feedback from previous trials influence current decisions? A study on the role of feedback processing in making decisions under explicit conditions. Journal of Neuropsychology, 2, 431-443.

Branscome, T. A., & Grynovicki, J. O. (2007). An investigation of factors affecting multi-task performance in an immersive environment (ARL Report No. TR-4325). Aberdeen, MD: U.S. Army Research Laboratory.

Bröder, A., & Schiffer, S. (2006). Adaptive flexibility and maladaptive routines in selecting fast and frugal decision strategies. Journal of Experimental Psychology: Learning, Memory, and Cognition, 32, 904–918.

Comstock, J. R., Jr., & Amegard, R. J. (1992). The Multi-Attribute Task Battery for human operator workload and strategic behavior research. National Aeronautics and Space Administration Technical Memorandum No. 104174.

Elsmore, T. F. (1994). SYNWORK 1: A PC-based tool for assessment of performance in a simulated work environment. Behavior Research Methods, Instruments, & Computers, 26, 421-426.

Lovett, M. C., & Anderson, J. R. (1996). History of success and current context in problem solving: Combined influences on operator selection. Cognitive Psychology, 31, 168-217.

Luchins, A. S. (1942). Mechanization in problem solving. Psychological Monographs, 54 (Whole no. 248).

Luwel, K., Schillemans, V., Onghena, P., & Verschaffel, L. (2009). Does switching between strategies within the same task involve a cost? British Journal of Psychology, 100, 753-771.

Mednick, M. T. (1963). Research creativity in psychology students. Journal of Consulting and Clinical Psychology, 27, 265-266.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on Mechanical Turk, Judgment and Decision Making, 5, 411-419.

Pinheiro, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-PLUS. New York: Springer Verlag.

Raven (1958). Raven's progressive matrices: a review and critical evaluation. Journal of General Psychology, 93, 199-228.

Reder, L. M. (1987). Strategy Selection in Question Answering. Cognitive Psychology, 19, 90-138.

Reder, L. M., & Schunn, C. D. (1999). Bringing together the psychometric and strategy worlds: Predicting adaptivity in a dynamic task. In Gopher, D. & Koriat, A. (Eds). Cognitive regulation of performance: Interaction of theory and application. Attention and Performance XVII (pp.315-342). Cambridge, MA: MIT Press.

Salthouse, T. A., & Mitchell, D. R. D. (1990). Effects of age and naturally occurring experience on spatial visualization performance. Developmental Psychology, 26, 845-854.

Shannon, C. E. (1948): A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379–423, 623–656.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. Science, 171, 701-703.

Shipley, W. C. (1946). Institute of living scale. Los Angeles: Western Psychological Services.

Tresselt, M. E., & Leeds, D. S. (1953). The Einstellung Effect in immediate and delayed problem-solving. Journal of General Psychology, 49, 87-95.
Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. Behavior Research Methods, 37, 498–505.
Wang, D. D., Proctor, R. W., & Pick, D. F. (2007). Acquisition and transfer of attention allocation strategies in a multiple-task work environment. Human Factors, 49, 995-1004.

APPENDIX 1

Detailed Description of RumRunner Parameters

RumRunner models the probability of strategy shift as the sum of factors:

Already Doing

Already Doing nullifies Task Shift in cases where the stimulus changes, but is now clearly biasing a strategy which the subject is already using. Like Task Shift, this factor is task-specific. In the drawing task, it is used for the Morph Conditions (Figure 1D) in Experiment 2, which progressively bias the subject towards a specific strategy. The value of this factor was between 0 and -0.1 in 71% of trials, and evenly distributed between -0.1 and -0.5 otherwise.

Bad Shift

Bad Shift doubles Performance Dip if the previous trial was a strategy shift.

Flexibility = $F_{-1}0.1(H_{-1} - P(H)_{-1})$

Flexibility is the base rate (i.e. prior probability) of strategy shift for the individual. It is initialized to 0 and updated according to the learning rate (0.1) and error term for the previous prediction, $H_{-1} - P(H)_{-1}$ where H_{-1} is 1 if strategy shifted on the previous trial and 0 otherwise, and $P(H)_{-1}$ is the output of RumRunner, i.e. the probability of shift on the previous trial.

No Improvement $= 1 - 1.1^{M}$

Where $M = \sum_{j=t-3}^{t-2}\sum_{k=j+1}^{t-1} S_k - S_j$ is the score slope, or average of the 3 most recent differences in successive scores, weighted by recency. It is only active when the slope is negative.
No Improvement is practically linear within the range of M observed in the experiment;
$f(M = -0.5) \cong 0.05;\ f(M = -0.01) \cong 0.00$

Number of Strategy Shifts $= 1.001^{-S} - 1$

S is the total number of strategy shifts by this individual (across all conditions). Number of Strategy Shifts is practically linear within the range of S observed in the experiment;
$f(S = 0) = 0;\ f(S = 20) \cong -0.02$

Performance Dip $= 1 - 1.1^{-C-1}$

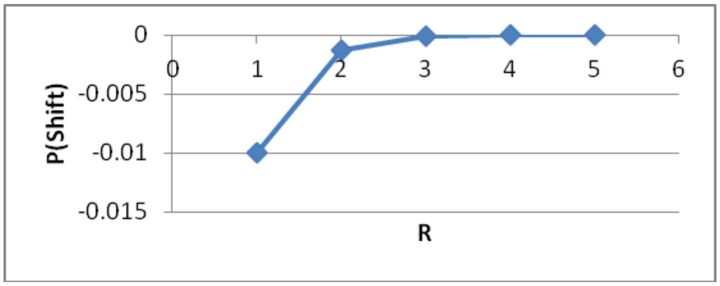
Where the decrease $C = {S_{min} - S_{t-1}}/{S_{max} - S_{min}}$, where S_{t-1} is the composite score on the previous trial and S_{min}, S_{max} are the minimum and maximum scores from the 10 preceding S_{t-1}.
Performance Dip is practically linear within the range of C observed in the experiment;
$f(C = 0) \cong 0.01;\ f(C = 1) \cong 0.17$

Planning/Preparation $= 1 - 1.1^{-D-1}$

Where the delay $D = {T - T_{max}}/{T_{max} - T_{min}}$, where T is response time, i.e. the number of seconds between presentation of stimulus and onset of drawing/tracing for the current trial. T_{max} and T_{min} are the maximum and minimum response times in the previous 10 trials.
Planning/Preparation is practically linear within the range of D observed in the experiment;
$f(D = 0) \cong 0.01;\ f(D = 5) \cong 0.44$

Recently Shifted $= -0.01R^{-R-1}$

Where R is the number of trials since the previous strategy shift.



Task Shift

This factor is task-dependent and reflects recent changes in stimulus that might prompt shift in strategy. In the drawing task, Task Shift is defined differently depending on how the stimulus is changing:

1) Degree of morph: the amount by which the current trial is “deformed” from the canonical shape (see Figure 1D; Experiment 2 only).
2) Degree of rotation: the amount by which the stimulus in the current trial is rotated, compared to the previous trial (e.g. difference in angles; Experiments 1 and 2).

The resulting distribution of Task Shift in the figure-drawing experiment approximates a one-sided normal distribution. Task Shift was between 0 and 0.1 in 38% of trials, between 0.4 and 0.5 in 10% of trials, and greater than 0.6 in 1% of trials.

Time on Task = $1 - 1.001^{-A-1}$

A is the total number of trials (across all conditions) completed by the individual. Time on Task is practically linear within the range of A observed in the experiments;

$$f(A = 0) \cong 0.00; \quad f(A = 80) \cong -0.08$$

DISTRIBUTION

1	Robert Abbott, 1463
1	James (Chris) Forsythe, 1462
1	Michael Haass, 1463
1	Stacey Hendrickson, 0413
1	Susan Stevens-Adams, 0413
1	RIM-Reports Management, 9532 (electronic copy)
1	MS0359 D. Chavez, LDRD Office 1911

Sandia National Laboratories